A Machine Learning Pipeline for estimating Binding Affinity to Serum Albumin and Half-Lives of Per- and Polyfluoroalkyl Substances in Humans

Haralambos Sarimveis[1], Vassilis Minadakis[1], Evie Papakyriakopoulou[1], Periklis Tsiros[1]
[1]National Technical University of Athens, School of Chemical engineering, 9 Heroon Polytechniou St, 15780, Athens, Greece

hsarimv@mail.ntua.gr

Per- and Polyfluoroalkyl substances (PFAS) are a broad group of chemicals, widely used in various applications, like aerospace and defense, electronics and textiles and coatings [1]. Their extensive production has raised significant concerns due to their adverse effects on human and animal health, as well as environmental systems [2]. A key concern regarding the use of PFAS is their long half-lives in both animals and humans. Especially for humans, the reported half-lives of many PFAS are in the range of years [3,4]. This persistence is attributed to their strong binding affinity to transporter proteins, particularly albumin, which is the most abundant protein in human and animal blood serum [5]. Understanding the binding behavior of PFAS to albumin and their resulting half-lives has driven the development of *in silico* tools to predict these critical endpoints [6,7].

Our work aims to develop a computational pipeline, capable of predicting the half-life of PFAS in humans, enhanced by predictions of their binding affinity to serum albumin. This workflow seeks to provide predictions across a diverse range of PFAS and identify factors that significantly influence their retention times in organisms. Specifically, we propose a pipeline that consists of two sequential machine learning models. The first model predicts the binding affinity (association constant) of PFAS to serum albumin. These predictions, along with additional features, serve as inputs to the second model, which estimates the half-life of PFAS. The proposed approach not only enables accurate half-life predictions but also facilitates exploration of the relationship between half-life and albumin binding affinity.

Both models belong to the quantitative structure-activity relationship (QSAR) class and make use of computational descriptors, which are computed from their simplified molecular-input line-entry system (SMILES) representations [8]. In addition to computational descriptors, experimental information such as the albumin assay type (e.g., equilibrium dialysis or fluorescence quenching) were included as features to account for variability in the datasets. Thus, separate training datasets were created for each QSAR model, containing computational descriptors and experimental data retrieved from the literature. These datasets were constructed to cover a wide range of PFAS and exposure scenarios, enhancing the models' extrapolation capabilities. To manage the high dimensionality of computational descriptors, feature selection techniques were applied. More specifically, features with low variance were removed and highly correlated features were eliminated to reduce redundancy.

To ensure reliable predictions, we defined the domain of applicability (DOA) of each QSAR model [9]. For PFAS congeners that are outside of the model's DOA, the corresponding prediction cannot be considered reliable. However, there is not yet a standard approach to define the DOA of a model. Multiple methodologies have been suggested in the literature and each one of them has its strengths and weak points [10]. In this work, we followed a strategy that implements multiple methodologies (Leverage, Mean-Variance, Bounding Box, Mahalanobis distance, Kernel-based and City block distance). The final decision about reliability of each prediction is derived from the majority voting of all methodologies. Robustness of the predictive QSAR models was examined through k-fold cross-validation, and Y-randomisation tests were performed to ensure that that the predictive ability of the models is not driven by chance correlations.

To gain insights about the relationships that exists between the endpoints and the independent features, we utilized SHapley Additive exPlanations (SHAP) values [11]. SHAP values provide a deep understanding of how molecular descriptors and experimental conditions influence binding affinity and half-life predictions. These values, grounded in game theory, are model-agnostic and offer multiple benefits: they identify outliers during training, provide explanations of how individual predictions are obtained by highlighting key contributing features and assess feature importance in the produced QSAR models [12].

The final stage of our work involves providing the QSAR models as online services. To achieve this, the models were deployed on Jaqpot (https://app.jaqpot.org), a self-developed online platform that hosts machine learning models and enables users to make predictions online. Deploying the models facilitates their integration into a streamlined pipeline that takes SMILES representations and other necessary data as input to predict the association constant of PFAS with albumin. The output can subsequently be used as input to the half-life QSAR model. The Jaqpot User Interface (UI) allows assessing the reliability of these predictions by verifying whether they fall within the models' DOA.

The developed models will be highly significant in the design of next-generation PFAS, aiming to enhance performance while addressing environmental concerns and minimizing adverse effects on humans and animals.

Acknowledgement

This work has been financially supported by the SCENARIOS project (Grant Agreement 101037509) which has been funded by the European Commission under the Horizon 2020 Programme.

References

[1] Glüge, J., Scheringer, M., Cousins, I.T., DeWitt, J.C., Goldenman, G., Herzke, D., Lohmann, R., Ng, C.A., Trier, X., Wang, Z., Environmental Science: Processes & Impacts, 22 (2020) 2345–2373.

[2] De Silva, A.O., Armitage, J.M., Bruton, T.A., Dassuncao, C., Heiger-Bernays, W., Hu, X.C., Kärrman, A., Kelly, B., Ng, C., Robuck, A., Sun, M., Webster, T.F., Sunderland, E.M., Environmental Toxicology and Chemistry, 40 (2021) 631–657.

[3] Li, Y., Fletcher, T., Mucs, D., et al., Occupational and Environmental Medicine, 75 (2018) 46–51.

[4] Olsen, G.W., Burris, J.M., Ehresman, D.J., Froehlich, J.W., Seacat, A.M., Butenhoff, J.L., Zobel, L.R., Environmental Health Perspectives, 115 (2007) 1298–1305.

[5] Han, X., Snow, T.A., Kemper, R.A., Jepson, G.W., Chemical Research in Toxicology, 16 (2003) 775–781.

[6] Gallagher, A., Kar, S., Sepúlveda, M.S., Molecules, 28 (2023) 5375.

[7] Dawson, D.E., Lau, C., Pradeep, P., Sayre, R.R., Judson, R.S., Tornero-Velez, R., Wambaugh, J.F., Toxics, 11 (2023) 98.

[8] O'Boyle, N.M., Journal of Cheminformatics, 4 (2012) 22.

[9] Weaver, S., Gleeson, M.P., Journal of Molecular Graphics and Modelling, 26 (2008) 1315–1326.

[10] Roy, K., Kar, S., Ambure, P., Chemometrics and Intelligent Laboratory Systems, 145 (2015) 22–29.

[11] Marcílio, W.E., Eler, D.M., 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Porto de Galinhas, Brazil, 2020, pp. 340-347.

[12] Rodríguez-Pérez, R., Bajorath, J., Journal of Computer-Aided Molecular Design, 34 (2020) 1013–1026.